Agreement of dermatopathologists in the evaluation of clinically difficult melanocytic lesions: how golden is the 'gold standard'?

Braun, R P ; Gutkowicz-Krusin, D ; Rabinovitz, H ; Cognetta, A ; Hofmann-Wellenhof, R ; Ahlgrimm-Siess, V ; Polsky, D ; Oliviero, M ; Kolm, I ; Googe, P ; King, R ; Prieto, V G ; French, L ; Marghoob, A ; Mihm, M

Abstract: BACKGROUND: The 'gold standard' for the diagnosis of melanocytic lesions is dermatopathology. Although most of the diagnostic criteria are clearly defined, the interpretation of histopathology slides may be subject to interobserver variability. OBJECTIVES: The aim of this study was to determine the variability among dermatopathologists in the interpretation of clinically difficult melanocytic lesions. METHODS: This study used the database of MelaFind®, a computer-vision system for the diagnosis of melanoma. All lesions were surgically removed and sent for independent evaluation by four dermatopathologists. Agreement was calculated using kappa statistics. RESULTS: A total of 1,249 pigmented melanocytic lesions were included. There was a substantial agreement among expert dermatopathologists: two-category kappa was 0.80 (melanoma vs. non-melanoma) and three-category kappa was 0.62 (malignant vs. borderline vs. benign melanocytic lesions). The agreement was significantly greater for patients 40 years (three-category kappa = 0.67) than for younger patients (kappa = 0.49). In addition, the agreement was significantly lower for patients with atypical mole syndrome (AMS) (kappa = 0.31) than for patients without AMS (kappa = 0.76). LIMITATIONS: The data were limited by the inclusion/exclusion criteria of the MelaFind® study. This might represent a selection bias. The agreement was evaluated using kappa statistics. This is a standard method for evaluating agreement among pathologists, but might be considered controversial by some statisticians. CONCLUSIONS: Expert dermatopathologists have a high level of agreement when diagnosing clinically difficult melanocytic lesions. However, even among expert dermatopathologists, the current 'gold standard' is not perfect. Our results indicate that lesions from younger patients and patients with AMS may be more problematic for the dermatopathologists, suggesting that improved diagnostic criteria are needed for such patients.

DOI: https://doi.org/10.1159/000336886

Original Paper

Dermatology

Dermatology 2012;224:51–58
DOI: 10.1159/000336886

Received: November 25, 2011
Accepted after revision: January 23, 2012
Published online: March 20, 2012

Agreement of Dermatopathologists in the Evaluation of Clinically Difficult Melanocytic Lesions: How Golden Is the 'Gold Standard'?

R.P. Braun[a] D. Gutkowicz-Krusin[b] H. Rabinovitz[c] A. Cognetta[d] R. Hofmann-Wellenhof[j] V. Ahlgrimm-Siess[j] D. Polsky[e] M. Oliviero[c] I. Kolm[a] P. Googe[g] R. King[g] V.G. Prieto[h] L. French[a] A. Marghoob[f] M. Mihm[i]

[a]Department of Dermatology, University Hospital Zürich, Zürich, Switzerland; [b]MELA Sciences, Inc., Irvington, N.Y., [c]Skin & Cancer Associates, Plantation, Fla., [d]Dermatology Associates of Tallahassee, Tallahassee, Fla., [e]Department of Dermatology, New York University School of Medicine, New York, N.Y., [f]Memorial Sloan-Kettering Cancer Center, New York, N.Y., [g]Knoxville Dermatopathology Laboratory, Knoxville, Tenn., [h]University of Texas M.D. Anderson Cancer Center, Houston, Tex., and [i]Department of Dermatology, Harvard Medical School, Boston, Mass., USA; [j]Department of Dermatology, Medical University, Graz, Austria

Key Words
Dermoscopy · Melanoma · Diagnosis · Histopathology · Agreement

Abstract
Background: The 'gold standard' for the diagnosis of melanocytic lesions is dermatopathology. Although most of the diagnostic criteria are clearly defined, the interpretation of histopathology slides may be subject to interobserver variability. ***Objectives:*** The aim of this study was to determine the variability among dermatopathologists in the interpretation of clinically difficult melanocytic lesions. ***Methods:*** This study used the database of MelaFind®, a computer-vision system for the diagnosis of melanoma. All lesions were surgically removed and sent for independent evaluation by four dermatopathologists. Agreement was calculated using kappa statistics. ***Results:*** A total of 1,249 pigmented melanocytic lesions were included. There was a substantial agreement among expert dermatopathologists: two-category kappa was 0.80 (melanoma vs. non-melanoma) and three-category kappa was 0.62 (malignant vs. borderline vs. benign melanocytic lesions). The agreement was significantly greater for patients ≥40 years (three-category kappa = 0.67) than for younger patients (kappa = 0.49). In addition, the agreement was significantly lower for patients with atypical mole syndrome (AMS) (kappa = 0.31) than for patients without AMS (kappa = 0.76). ***Limitations:*** The data were limited by the inclusion/exclusion criteria of the MelaFind® study. This might represent a selection bias. The agreement was evaluated using kappa statistics. This is a standard method for evaluating agreement among pathologists, but might be considered controversial by some statisticians. ***Conclusions:*** Expert dermatopathologists have a high level of agreement when diagnosing clinically difficult melanocytic lesions. However, even among expert dermatopathologists, the current 'gold standard' is not perfect. Our results indicate that lesions from younger patients and patients with AMS may be more problematic for the dermatopathologists, suggesting that improved diagnostic criteria are needed for such patients.

Parts of this work were presented at the 2009 Annual Meeting of the Swiss Society of Dermatology and Venereology in Basel, Switzerland.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1018–8665/12/2241–0051$38.00/0

Accessible online at:
www.karger.com/drm

R.P. Braun
Department of Dermatology, University Hospital Zürich
Gloriastrasse 31
CH–8091 Zürich (Switzerland)
Tel. +41 44 255 9481, E-Mail ralph.braun@usz.ch

Introduction

Accurate interpretation of the histopathology of biopsied specimens is essential in directing appropriate patient management, that is how the patients are informed, treated and followed up. In addition, the final pathology diagnosis, whether correct or not, is ultimately responsible for determining which lesions enter into cancer registry databases [1]. This in turn may influence the results of epidemiological investigations of cancer, diagnostic studies, or comparative studies that rely upon evaluations provided by pathologists [2–5].

Evidence-based medicine requires that the performance of a diagnostic test be compared to a reference standard (i.e. 'gold standard') [6]. In dermatology, clinical studies designed to evaluate the performance of diagnostic tests, such as clinical examination, dermoscopic examination, or computer-assisted diagnosis, require that the results be compared to a reference standard. This 'reference' or 'gold' standard is currently the dermatopathological diagnosis [2, 7, 8], based mainly on visual interpretation of formalin-fixed histological slides of the biopsied tissue. Although most of the diagnostic criteria for differentiating melanocytic lesions are clearly defined, the interpretation of the histopathology slides may be subject to interobserver variability since it relies on subjective evaluation of tissue morphology and of cytological atypia. The aim of this study was to determine the degree of interobserver variability among dermatopathologists in the interpretation of clinically difficult to diagnose melanocytic lesions of the skin.

Materials and Methods

This study utilized the MelaFind® database. MelaFind® is a computer-vision system designed by MELA Sciences, Inc. (MELA) to identify lesions to be biopsied to rule out melanoma; clinical testing of this system has been recently completed [9, 10]. Data for this study were collected prospectively at multiple clinical sites in the U.S., Switzerland, Austria, and Australia. According to the MelaFind® study protocol, all patients (without age limit) scheduled for the surgical removal of any pigmented skin lesion which met the inclusion criteria were eligible to participate in this study. The inclusion criteria required that the pigmented lesion (melanocytic or non-melanocytic) be between 2 and 22 mm in diameter and be devoid of ulceration, bleeding, or foreign matter (e.g. tattoos). Furthermore, lesions on mucosa or in subungual regions were excluded.

For each case, the following clinical information was recorded: patient age, gender, anatomic location of the lesion, and pre-biopsy clinical and dermoscopic diagnosis rendered by the examining dermatologist at the time of examination. The clinician was requested to select from the following diagnostic possibilities: 'melanoma', 'melanoma cannot be ruled out', or 'not melanoma'. If the lesion was not felt to be melanoma but still removed, the clinician was asked to record the reason for biopsy. The lesions were surgically removed and subjected to routine histopathology processing at the treating physician's preferred dermatopathology laboratory. The diagnostic histological slides were subsequently sent to MELA for independent histopathological evaluation by a panel of four dermatopathologists. The data from five clinical sites (AC, RB, RHW, DP, HR) were used in the present study. Since our focus was the diagnosis of melanocytic lesions, all other lesions (e.g. basal cell carcinoma, pigmented seborrheic keratosis, etc.) were excluded from this study, based on the final histological interpretation. Each histopathology case was examined independently by two study dermatopathologists out of a panel of four experts (M.M., P.G., R.K., V.G.P.). The diagnosis of the local pathologist was not recorded or taken into account for the purpose of this study and only the diagnoses of the study pathologists were recorded and used in the development of the computer-vision system.

The four study dermatopathologists did not partake in any formal or informal training or consensus meeting at any time for the purpose of this study. However, it should be acknowledged that two of the study dermatopathologists were trained by Dr. Mihm (Dr. Googe about 20 years ago and Dr. King about 10 years ago), and that Dr. Mihm evaluated all cases while each of the remaining study dermatopathologists evaluated about 1/3 of the cases.

The histopathological diagnosis was ultimately used as the reference standard for the evaluation of the diagnostic performance of the computer-vision system [11]. In case of significant discordance between the first two dermatopathologists, the histological slide was independently reviewed by a third study pathologist. According to our definition 'significant discordance' occurred if one study pathologist read a slide as MM or HGDN (positive diagnosis) and the second pathologist read it as LGDN or OTHER (negative diagnosis). The dermatopathologists were completely blinded to the pre-biopsy diagnosis, originating laboratory pathology diagnosis, patient demographics such as age, or diagnosis made by other study pathologists. In addition, the pathologists were unaware whether the case was being reviewed as a regular evaluation or due to discordance between the first two study pathologists.

Interobserver agreement was evaluated using the kappa statistic and kappa values were compared using a χ^2 test, as defined by Fleiss [12]. For interpretation of the kappa values, we used the sys-

Table 1. Interpretation of kappa values according to Landis and Koch [13]

Kappa	Interpretation
<0	no agreement
0.00–0.20	slight agreement
0.21–0.40	fair agreement
0.41–0.60	moderate agreement
0.61–0.80	substantial agreement
0.81–1.00	almost perfect agreement

tem proposed by Landis and Koch [13] (table 1). Weighted kappa values were used to take into account the fact that discordance of diagnoses such as melanoma versus junctional nevus is clinically much more important than the discordance of diagnoses such as melanoma in situ versus HGDN or HGDN versus LGDN. In addition, such values allow a comparison of our results with the results previously reported in the literature that were based on the weighted kappa statistics. While kappa statistics have been criticized in the literature [14, 15], they are recommended for the study of observer agreement by others [16, 17].

Results

A total of 1,249 melanocytic pigmented lesions from five clinical sites matched the inclusion criteria for this study (table 2). The final histological diagnoses were established as follows. A lesion was considered (1) melanoma if at least one (blindly verified) diagnosis was melanoma, (2) HGDN if at least two diagnoses were HGDN, or (3) OTHER in the remaining cases. Only 190 cases (15.2%) were considered 'not melanoma' and 97 cases (7.8%) were considered 'melanoma' prior to biopsy by the examining dermatologists. The remaining lesions were biopsied due to suspicion of melanoma (i.e. 'melanoma cannot be ruled out'). Thus, almost 80% of lesions included in the study were difficult to diagnose both clinically and dermoscopically. The percentages of histologically confirmed melanomas among lesions with the clinical diagnoses of 'melanoma', 'melanoma cannot be ruled out', and 'not melanoma' were 82.5, 11.5, and 2.6%, respectively.

Of the 1,249 lesions, 129 (10.3%) required an evaluation by a third study panel dermatopathologist because of significant discordance, as defined above. All other cases had concordant histopathological diagnoses indicating a good level of agreement among dermatopathologists. Kappa statistics were utilized to analyze agreement, and since results depend on the number of categories and the prevalence of these categories in the database, we evaluated kappa for several types of stratification of the data. Since ten different comparisons were made in this study, we adopted the conservative Bonferroni method so that each test is considered statistically significant if $p < 0.005$ [17].

First, we calculated kappa values for two categories: 'melanoma' (invasive or in situ) and 'non-melanoma' (HGDN/AMP/AMH, LGDN, and other benign nevi). The corresponding kappa values for these two categories are shown in table 3. We found excellent agreement among dermatopathologists for the diagnosis of melanoma versus non-melanoma, with an overall kappa value of 0.80. The agreement among dermatopathologists was found to be greater if the patient's age was ≥40 years ($p = 0.03$, but not statistically significant for multiple comparisons).

Next, we determined the agreement between dermatopathologists using three categories: 'MALIGNANT' (in situ melanoma, invasive melanoma), 'UNCERTAIN' (HGDN/AMP/AMH), and 'BENIGN' lesions (all other melanocytic neoplasms). The corresponding kappa values are shown in table 4. The overall kappa value was 0.62, indicating substantial agreement. As expected, an increase in the number of categories leads to a decrease in kappa values.

Although agreement was slightly better for lesions with a definite pre-biopsy diagnosis as rendered by the

Table 2. Summary of the final histological diagnoses in the database

Lesion type	Cases	Prevalence
Melanoma, invasive	116	9.3%
BT median = 0.44 mm		
BT range 0.15–5.00 mm		
Melanoma, in situ	80	6.4%
Nevus, high-grade dysplastic/AMP/AMH	75	6.0%
Nevus, low-grade dysplastic	747	59.8%
Nevus, blue	20	1.6%
Nevus, congenital	46	3.7%
Nevus, Spitz	10	0.8%
Nevus, other	155	12.4%
Total	1,249	100.0%

AMH = Atypical melanocytic hyperplasia; AMP = atypical melanocytic proliferation; BT = Breslow thickness.

Table 3. Kappa statistics for two categories: 'melanoma' (invasive or in situ) and 'non-melanoma' (HGDN/AMP/AMH, LGDN, and other benign nevi)

Type	Subtype	Cases	Kappa (95% CI)	p
All		1,249	0.80 (0.75–0.86)	
Patient age, years	1–40	594	0.69 (0.61–0.77)	0.03
	>40	655	0.82 (0.74–0.90)	
Patient sex	female	592	0.84 (0.76–0.92)	0.27
	male	657	0.77 (0.70–0.85)	
Pre-biopsy diagnosis	definite	287	0.83 (0.72–0.95)	0.34
	suspicious	962	0.77 (0.70–0.83)	

Table 4. Kappa statistics for three categories: 'MALIGNANT' (in situ and invasive melanoma), 'UNCERTAIN' (HGDN/AMP/AMH), and 'BENIGN' (all other benign melanocytic neoplasms)

Type	Subtype	Cases	Kappa (95% CI)	p
All		1,249	0.62 (0.58–0.67)	
Patient age, years	1–40	594	0.49 (0.43–0.55)	<0.0001
	>40	655	0.67 (0.61–0.73)	
Patient sex	female	592	0.67 (0.61–0.73)	0.06
	male	657	0.59 (0.53–0.65)	
Pre-biopsy diagnosis	definite	287	0.68 (0.59–0.78)	0.07
	suspicious	962	0.59 (0.54–0.65)	

Table 5. Kappa statistics for three categories (MM, HGDN/AMP/AMH, and OTHER) for patients with and without AMS

AMS status	Cases	Kappa	p
With AMS	212	0.31	<0.0001
Without AMS	146	0.76	
With AMS ≤40 years	155	0.07	0.001
With AMS >40 years	57	0.48	
Without AMS ≤40 years	50	0.56	0.12
Without AMS >40 years	96	0.79	

examining dermatologist (melanoma or not melanoma) than for lesions having an unclear pre-biopsy diagnosis ('melanoma cannot be ruled out'), this did not meet statistical significance.

Once again, the overall agreement was greater for patients ≥40 years (kappa = 0.67) as compared to younger patients (kappa = 0.49), and this was statistically highly significant ($p < 0.0005$). Among younger patients, there were 111 (19%) lesions with at least one positive (melanoma or HGDN) pathological diagnosis and 75 (68%) of these lesions had discordant diagnoses. Among older patients, there were 229 (35%) lesions with at least one positive diagnosis and 98 (43%) of these lesions had discordant diagnoses. Thus, the kappa statistic does differentiate between patients with low and high agreement of pathological diagnoses.

Since actinic damage of the skin may influence the interpretation of the histopathology, we compared the data from different clinical centers separately. The data showed no significant difference in variability of diagnosis between clinical centers having many patients with sun-damaged skin as compared to clinical centers with few such patients: for the two Florida sites (538 cases), the two-category kappa (melanoma vs. not melanoma) was 0.79 while for the other three sites (711 cases) it was 0.80. Thus, our data showed no significant difference in the variability of diagnosis between sites with many patients having sun-damaged skin compared to sites with few such patients.

In addition, we also retrospectively evaluated all cases submitted by HR and RB to determine whether the patients had the atypical mole syndrome (AMS) or not. We used the definition of AMS given by Kopf et al. [18, 19]: 'Patients with 100 or more melanocytic nevi, one or more melanocytic nevi at least 8 mm in maximum diameter, and one or more nevi with clinical atypical features.' Overall there were 212 cases with AMS and 146 cases without AMS at these two sites; for some patients the AMS status was not known and these cases were excluded from the analysis. The corresponding kappa values are shown in table 5. The kappa values indicate that pathological agreement was substantially lower for specimens derived from patients with AMS (kappa = 0.31) as compared to patients without AMS (kappa = 0.76); the difference is highly significant ($p < 0.001$).

In the corresponding subgroup analysis, we found that for patients with AMS, the agreement was still significantly lower for the younger patients (<40 years). For patients without AMS the difference in pathological agreement was not significantly different for younger and older patients. This could be partially due to the small sample size (50 patients <40 years, 96 patients ≥40 years). Our results suggest that the agreement among dermatopathologists depends on both patient's age and patient's AMS status.

Lastly, we analyzed agreement for five categories of histopathological diagnosis: MM invasive, MM in situ, HGDN/AMP/AMH, LGDN, and OTHER. The results are shown in table 6, where prevalence denotes frequency of the diagnosis rather than final histological diagnosis. We found the highest agreement for invasive melanoma (kappa = 0.74) and second highest for melanoma in situ (kappa = 0.54). The agreement was lowest for high-grade dysplastic nevi and it was slightly higher for low-grade dysplastic nevi. The overall kappa values indicate moderate agreement.

In statistics 'weighting' is used to account for the fact that not all disagreements are of equal clinical importance; for example, a disagreement on two diagnoses such as melanoma vs. junctional nevus would be clinically much more relevant than the disagreement of diagnoses

such as melanoma in situ vs. HGDN or HGDN vs. LGDN. In our analysis we used two different types of weighting. Both Cohen's and Cichetti-Allison's weighting formulas [12] indicate substantial agreement between the study dermatopathologists.

In addition to considering invasive versus in situ melanomas, it would also be of interest to consider thick melanomas (>1 mm in thickness) vs. thin melanomas (in situ and ≤1 mm in thickness). However, the data set of 1,249 melanocytic lesions includes only 17 lesions with at least one histological diagnosis of thick melanoma and this sample is too small for kappa statistics. Furthermore, two of these cases have only one determination of thickness. For the remaining cases of thick melanoma, the concordance rate was 80% (12 out of 15.) Among 179 lesions with at least one histological diagnosis of thin melanoma, the concordance rate was 70% (125 out of 179.) The rate of concordance was not significantly different for thin or thick melanomas, with $p = 0.56$ by Fisher's exact test.

Table 6. Kappa statistics for five categories: MM invasive, MM in situ, HGDN/AMP/AMH, LGDN, and OTHER

Diagnosis category	Prevalence	Kappa (95% CI)
MM invasive	7.2%	0.74 (0.68–0.79)
MM in situ	5.5%	0.54 (0.49–0.60)
HGDN/AMP	9.4%	0.24 (0.18–0.29)
LGDN	47.7%	0.32 (0.27–0.38)
OTHER	30.2%	0.38 (0.33–0.44)
OVERALL, unweighted		0.39 (0.36–0.42)
OVERALL, weight w_1		0.79 (0.72–0.85)
OVERALL, weight w_2		0.61 (0.56–0.66)

Weights w_1 and w_2 are according to Cohen and Cicchetti-Allison, respectively.

Discussion

Evidence-based medicine requires that new diagnostic tests must be compared to the currently accepted reference standard (i.e. 'gold standard') in order to determine the new test's performance and merit. Therefore, new diagnostic techniques utilized for the diagnosis of melanocytic neoplasms, such as dermoscopy, confocal microscopy, and machine-vision, just to mention a few, have generally been compared to dermatopathology, which is currently considered the gold standard. We are convinced that the 'true gold standard' in cancer diagnosis is not a diagnostic test but the malignant behavior of the neoplasm. Even if a gold standard classifies a lesion as benign, the occurrence of metastatic behavior will prove the gold standard wrong. However, even though the diagnostic criteria for melanoma are well established in dermatopathology, it is important to acknowledge that this gold standard is based on visual examination of tissue morphology and cytological atypia, which, like any other visual evaluation, is open to subjective interpretation. Dermatopathology is subject to interobserver variability. With that being said, one of the quandaries in evidence-based medicine is determining how and when it is permissible to overrule the gold standard or when is it time to replace the gold standard with a better one. In other words, if a new diagnostic test outperforms the current gold standard, how is it possible to prove that the new test is in fact superior to the current standard? For example, let us assume that the lesion under investigation is biologically a melanoma. If a diagnostic method renders a diagnosis of melanoma and the dermatopathologist renders a diagnosis of benign nevus, evidence-based medicine may classify the lesion as benign, when in fact it is melanoma, and the result of the new instrument will be considered false positive. Rarely is consideration given to the fact that a diagnostic method may potentially be superior to the current gold standard.

Since more and more non-invasive technologies and new in vivo and ex vivo tests for melanoma are emerging, we felt that there was a need to take a closer look at what is currently considered to be the gold standard for the diagnosis of melanoma. We therefore investigated the degree of interobserver agreement for the histopathological diagnosis of difficult melanocytic neoplasms by expert dermatopathologists. We used the MelaFind® database, since the data have been collected prospectively in different centers under routine clinical conditions, and the histopathological evaluation was performed by a panel of dermatopathologists. This setting provided data that were ideal for the purpose of this study.

Since our focus was the diagnosis of melanocytic lesions, we excluded all other lesions (e.g. basal cell carcinoma, pigmented seborrheic keratosis, etc.). We included the data from five clinical sites, which together contributed most of the cases in the MelaFind® database. The clinical sites were either specialized referral centers or pigmented skin lesion clinics (RB, DP, RHW) or private practices specializing in the diagnosis and treatment of skin cancers (AC, HR). All five centers see mainly high-risk melanoma patients such as those with AMS, previous his-

tory of skin cancer, family history of melanoma, etc. From our own experience, we are of the opinion that the pigmented lesions in many of these high-risk patients are more difficult to diagnose because many of their benign nevi manifest clinical and/or dermoscopic features of melanoma or their melanomas manifest subtle features that make their detection a challenge. Cuellar et al. recently reported that in fair-skinned individuals early melanomas can be difficult to diagnose [20]. Thus, the MelaFind® database is biased towards lesions with a higher degree of diagnostic difficulty as compared to lesions encountered in the setting of a general dermatology office.

To our knowledge, this is the first large-scale study (n = 1,249 lesions) of the agreement among dermatopathologists for the diagnosis of clinically difficult melanocytic neoplasms. Even though the lesions were of a high degree of diagnostic difficulty, only 10.3% required a third evaluation by a study dermatopathologist. The interobserver agreement among expert dermatopathologists was found to be better than reported previously in the literature, and we found that the results of the literature were contradictory, most likely due to different selection biases in different studies.

In a recent retrospective study, Shoo et al. found discordant diagnoses in 14.3% of 392 cases referred to a specialized center [21]. This might represent a selection bias because it is likely that only the most difficult cases are referred for second opinion. As a consequence, the agreement will be low due to the selection of difficult lesions. In order to avoid this potential bias we included every biopsied lesion from the study sites. In addition, our study pathologists were all expert dermatopathologists specialized in the diagnosis of melanocytic neoplasms. This explains the higher agreement rate in our study.

In another study reported by Farmer et al., 37 histological slides were evaluated by eight pathologists and the results recorded in three categories (benign, malignant, and indeterminate) [22]. They found a combined kappa of 0.50, as compared to 0.62 in our study.

In a publication by Duncan et al., the interobserver agreement based on 60 melanocytic lesions revealed kappa values between 0.55 and 0.84 for the analysis in two categories (melanoma vs. not melanoma) [23].

Another study by Corona et al. used 140 slides (mainly melanoma with a small subset of benign pigmented lesions) and found a kappa of 0.61 [24] as compared to 0.80 in our study. In conclusion, the authors found 'considerable disagreement among pathologists on the diagnosis of melanoma versus other pigmented lesions'.

Cook et al. reported the results of a study of 95 histological slides of thin melanomas that had been read by eight pathologists. The two-category kappa in this study was 0.77 [25].

Ferrara et al. investigated interobserver agreement based on 107 cases of dermoscopically and histopathologically equivocal melanocytic lesions and found an overall kappa of 0.74. In their initial pilot study from 36 possible pairs of observers, 8 pairs showed a kappa >0.75; 19 showed a kappa >0.5 and 17 had a kappa <0.5, indicating fair to moderate agreement [26, 27].

In the present study, based on the analysis of 1,249 lesions, the overall kappa was 0.80. It is important to mention that there was no initial meeting or consensus conference between the study dermatopathologists and that they were totally blinded to the clinical diagnosis, level of clinical suspicion, or patient demographics while evaluating the pathology slides. Furthermore, there was absolutely no exchange of information between the study dermatopathologists regarding the study. This is an important point since such communications could have created a consensus among the pathologists on how to interpret and diagnose the pathology slides, thereby leading to an artificially high concordance. With that being said, we do need to acknowledge that some of the study pathologists had worked together previously and this may have partially contributed to the high agreement rate. The alternative, and perhaps more likely, explanation as to why the agreement in our study was high is that all the study dermatopathologists are experts specializing in the diagnosis of skin cancers.

An important point is the nature and number of the different diagnostic categories used for the classification of melanocytic neoplasms. According to different schools and concepts in histopathology there are different types of categories. This issue has been addressed by Zembowicz and Scolyer [28] in a recent review article. The authors came to the conclusion that it was more appropriate to expand the classification scheme of melanocytic neoplasms rather than to narrow it. This strongly supports the classification used in this study (table 2).

Considering different diagnostic categories, we found that agreement was highest for the diagnosis of invasive melanoma (kappa = 0.74), second highest for melanoma in situ (kappa = 0.54) and worst for high-grade dysplastic nevi (kappa = 0.24) (table 6). This confirms the observation made by Duncan et al. [23] that the histopathology concordance for the different grades of dysplastic nevi was rather poor. The agreement was very good for the majority of cases and rather poor in the remaining chal-

lenging cases. Our findings suggest that the category of high-grade dysplastic nevi had the largest number of challenging cases.

The agreement of the dermatopathologists was better if the clinician was sure of his clinical diagnosis (either melanoma or non-melanoma). If the pre-biopsy diagnosis was 'melanoma cannot be ruled out', indicating that the clinician was not sure whether the lesion was in fact benign or not, the agreement among the pathologists was poorer, but the difference was not statistically significant. Ferrara et al. looked at the concordance of dermoscopy and histopathology and came to similar conclusions [26, 27]. If different dermoscopists agreed on the dermoscopic diagnosis, different dermatopathologists also tended to agree on the histopathology diagnosis and vice versa.

To our knowledge, none of the previous studies looked at other parameters such as patient age or gender. It is interesting that the agreement of the dermatopathological diagnosis was significantly higher in patients ≥40 years. Our first hypothesis was that with increasing age, patients might have more cumulative sun exposure and sun damage and that the sun damage may somehow account for the higher concordance. Two of the study sites are located in Florida (AC, HR) where even younger patients (<40 years) already manifest extensive sun damage. On the other hand, the other study sites (RB, RHW and DP) are located in geographic areas where patients have much less sun exposure and manifest less sun damage. There was no significant difference between agreement at the sites in different geographical areas.

Our next hypothesis was that most of the younger patients seen in these specialized referral centers were at high risk for developing melanoma and that a significant proportion of them may manifest the AMS. Based on this premise, at two study sites (RB and HR), the principal investigators went back to the charts of all patients enrolled in this study and recorded if the patients had AMS or not; patients with unknown AMS status were excluded from the analysis. We compared the pathological kappa values for patients with (n = 212) and without (n = 146) AMS and found that kappa was significantly lower for patients with AMS (kappa = 0.31) as compared to patients without AMS (kappa = 0.76) ($p < 0.0001$). Thus, our data indicate that lesions from patients with AMS may be more problematic for the pathologists, accounting for the lower agreement among the dermatopathologists. However, within the subgroup of AMS patients, the agreement was still significantly lower for the younger patients.

In our opinion, these findings could be explained by timing of nevogenesis and senescence [29–33]. The process of nevogenesis in non-AMS individuals appears to be more frequent and more dynamic in youth. By the time individuals are in their fourth decade of life, most nevi have entered into a senescent state and the frequency of developing new or enlarging nevi becomes much lower [29, 34, 35]. In contrast, the nevogenesis process may be more exuberant in individuals with AMS. Thus, in addition to developing numerous new nevi in adolescence and youth, many patients with AMS continue to develop new and growing nevi in older age. These new nevi, which have not yet undergone senescence, may manifest a variety of atypical morphologies, which in turn may explain the observed poorer agreement among the study pathologists [29]. Although this remains a conjecture, our study clearly indicates that agreement among pathologists was poorer in patients <40 years and in patients manifesting AMS.

Conclusions

Expert dermatopathologists have a high level of agreement when diagnosing clinically difficult melanocytic lesions. However, even though this agreement is very good, it is not 100%. Thus, even among expert dermatopathologists, the current 'gold standard' is not perfect and this fact needs to be addressed when deciding on patient management or when evaluating new diagnostic methods. Furthermore, this study highlights the fact that any discipline that relies on the visual evaluation of tissue morphology for rendering a diagnosis, whether clinical, dermoscopic, radiological or histological, must contend with interobserver disagreements, due to differences of perception or differences in interpretation. Therefore, clinicians must reconcile the clinical and histological information, obtain second opinions when appropriate, and ultimately assimilate all the information in a manner that allows for the formulation of a reasonable management plan that keeps the best interest of the patient foremost. Nevertheless, until proven otherwise, histopathology remains the best method available at this time for rendering an accurate diagnosis of melanocytic neoplasms.

While agreement of histological diagnoses by expert dermatopathologists is very good, it is significantly poorer for younger patients and for patients with AMS. Therefore, the results of this study suggest that improved diagnostic criteria are needed for these patients.

Acknowledgements

The authors thank MELA Sciences, Inc. for releasing data for this study, and Joanna Adrian, Mrinalini Roy, and Nikolai Kabelev of MELA for help with data management. The work of Dr. Braun and Dr. Kolm was supported by an Oncosuisse grant as well as by the Swiss National Foundation. The authors declare no other funding.

Disclosure Statement

D.G.K. is an employee of MELA Sciences, Inc.; M.M., V.G.P., P.G. and R.K. were the dermatopathologists for the MelaFind® study; R.P.B., I.K., H.R., M.O., A.C., D.P., R.H.W. and V.A.S. were investigators in the MelaFind® study. There are no conflicts of interest relevant for this study.

References

1 Thompson B, Austin R, Coory M, Aitken JF, Walpole E, Francis G, Fritschi L: Completeness of histopathology reporting of melanoma in a high-incidence geographical region. Dermatology 2009;218:7–14.
2 Mayer J: Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. Med J Aust 1997;167: 206–210.
3 Lock-Andersen J, Hou-Jensen K, Hansen J, Jensen N, Sogaard H, Andersen P: Observer variation in histological classification of cutaneous malignant melanoma. Scand J Plast Reconstr Surg Hand Surg 1995;29:141–148.
4 Busam K, Antonescu C, Marghoob M, Nehal K, Sachs D, Shia J, Berwick M: Histologic classification of tumor-infiltrating lymphocytes in primary cutaneous malignant melanoma. A study of interobserver agreement. Am J Clin Pathol 2001;115:856–860.
5 Oliveria S, Dusza S, Berwick M: Issues in the epidemiology of melanoma. Expert Rev Anticancer Ther 2001;1:453–459.
6 Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB: Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh/New York, Churchill Livingstone, 2000.
7 Kittler H, Pehamberger H, Wolff K, Binder M: Diagnostic accuracy of dermoscopy. Lancet Oncol 2002;3:159–165.
8 Vestergaard M, Macaskill P, Holt P, Menzies SW: Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br J Dermatol 2008;159:669–676.
9 Elbaum M, Kopf AW, Rabinovitz HS, et al: Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: a feasibility study. J Am Acad Dermatol 2001;44:207–218.
10 Friedman RJ, Gutkowicz-Krusin D, Farber MJ, et al: The diagnostic performance of expert dermoscopists vs a computer-vision system on small-diameter melanomas. Arch Dermatol 2008;144:476–482.
11 Monheit G, Cognetta AB, Ferris L, et al: The performance of MelaFind: a prospective multicenter study. Arch Dermatol 2011; 147:188–194.
12 Fleiss JL: Statistical Methods for Rates and Proportions. New York, John Wiley, 1981.
13 Landis JR, Koch GG: The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.
14 Uebersax JS: Diversity of decision-making models and the measurement of interrater agreement. Psychol Bull 1987;101:140–146.
15 Uebersax JS: Modeling approaches for the analysis of observer agreement. Invest Radiol 1992;27:738–743.
16 Kundel HL, Polansky M: Measurement of observer agreement. Radiology 2003;228: 303–308.
17 Fisher LD: Biostatistics. New York, Wiley, 1993.
18 Marghoob AA, Kopf AW, Rigel DS, et al: Risk of cutaneous malignant melanoma in patients with 'classic' atypical-mole syndrome. A case-control study. Arch Dermatol 1994;130:993–998.
19 Kopf AW, Friedman RJ, Rigel DS: Atypical mole syndrome. J Am Acad Dermatol 1990; 22:117–118.
20 Cuellar F, Puig S, Kolm I, Puig-Butille S, Zaballos P, Marti-Laborda R, Badenas C, Malvehy J: Dermoscopic features of melanomas associated with MC1R variants in Spanish CDKN2A mutation carriers. Br J Dermatol 2009;160:48–53.
21 Shoo BA, Sagebiel RW, Kashani-Sabet M: Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center. J Am Acad Dermatol 2010;62:751–756.
22 Farmer ER, Gonin R, Hanna MP: Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. Hum Pathol 1996;27:528–531.
23 Duncan L, Berwick M, Bruijn J, Byers H, Mihm M, Barnhill R: Histopathologic recognition and grading of dysplastic melanocytic nevi: an interobserver agreement study. J Invest Dermatol 1993;100:318S–321S.
24 Corona R, Mele A, Amini M, et al: Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. J Clin Oncol 1996; 14:1218–1223.
25 Cook M, Clarke T, Humphreys S, et al: The evaluation of diagnostic and prognostic criteria and the terminology of thin cutaneous malignant melanoma by the CRC Melanoma Pathology Panel. Histopathology 1996;28: 497–512.
26 Ferrara G, Argenziano G, Soyer HP, et al: Dermoscopic and histopathologic diagnosis of equivocal melanocytic skin lesions: an interdisciplinary study on 107 cases. Cancer 2002;95:1094–1100.
27 Ferrara G, Argenziano G, Soyer HP, et al: Histopathologic interobserver agreement on the diagnosis of melanocytic skin lesions with equivocal dermoscopic features: a pilot study. Tumori 2000;86:445–449.
28 Zembowicz A, Scolyer RA: Nevus/melanocytoma/melanoma: an emerging paradigm for classification of melanocytic neoplasms? Arch Pathol Lab Med 2011;135:300–306.
29 Grichnik J: Melanoma, nevogenesis, and stem cell biology. J Invest Dermatol 2008; 128:2365–2380.
30 Zalaudek I, Hofmann-Wellenhof R, Soyer HP, Ferrara G, Argenziano G: Naevogenesis: new thoughts based on dermoscopy. Br J Dermatol 2006;154:793–794.
31 Zalaudek I, Ferrara G, Argenziano G: Dermoscopy insights into nevogenesis: 'Abtropfung' versus 'Hochsteigerung'. Arch Dermatol 2007;143:284.
32 Zalaudek I, Marghoob AA, Scope A, Hofmann-Wellenhof R, Ferrara G, Argenziano G: Age distribution of biopsied junctional nevi – Unna's concept versus a dual concept of nevogenesis. J Am Acad Dermatol 2007; 57:1096–1097.
33 Zalaudek I, Hofmann-Wellenhof R, Kittler H, et al: A dual concept of nevogenesis: theoretical considerations based on dermoscopic features of melanocytic nevi. J Dtsch Dermatol Ges 2007;5:985–992.
34 Aguilera P, Puig S: Clinical and dermoscopic features of nevi in children. Dermatology 2010;220:54.
35 Barnhill RL: Melanocytic nevi and tumor progression: perspectives concerning histomorphology, melanoma risk and molecular genetics. Dermatology 1993;187:86–90.